QMRF identifier (JRC Inventory): **To be entered by JRC**

QMRF Title: **IRFMN/ISSCAN-CGX expert rule-based model for carcinogenicity (v. 1.0.2)**

Printing Date: June 6, 2022

1.QSAR identifier

1.1.QSAR identifier (title):

IRFMN/ISSCAN-CGX expert rule-based model for carcinogenicity (v 1.0.2)

1.2.Other related models:

NA

1.3.Software coding the model:

VEGA (https://www.vegahub.eu/)

The VEGA software provides QSAR models to predict tox, ecotox, environ, phys-chem and toxicokinetic properties of chemical substances.

emilio.benfenati@marionegri.it

2.General information

2.1.Date of QMRF:

June 2022

2.2.QMRF author(s) and contact details:

[1] Simona Kovarich S-In Soluzioni Informatiche Soluzioni Informatiche Srl Via Ferrari 14, I-36100Vicenza simona.kovarich@s-in.it http://www.s-in.it https://www.marionegri.it/

[2] Erika Colombo – IRCCS Istituto di Ricerche Farmacologiche Mario Negri – erika.colombo@marionegri.it

2.3.Date of QMRF update(s):

NA

2.4.QMRF update(s):

NA

2.5.Model developer(s) and contact details:

[1] Alberto Manganaro Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2,20156 Milano, Italy alberto.manganaro@marionegri.it https://www.marionegri.it/

[2] Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2,20156 Milano, Italy emilio.benfenati@marionegri.it https://www.marionegri.it/

2.6.Date of model development and/or publication:

2016

2.7.Reference(s) to main scientific papers and/or software package:

[1] A. Golbamaki, E. Benfenati, N. Golbamaki, A. Manganaro, E. Merdivan, A. Roncaglioni, G. Gini(2016) New clues on carcinogenicity-related substructures derived from mining two large datasets of chemical compounds. JOURNAL OF ENVIRONMENTAL SCIENCE AND HEALTH, PART C, VOL.34, NO. 2, 97-113 http://dx.doi.org/10.1080/10590501.2016.1166879

[2] Benfenati E, Manganaro A, Gini G. VEGA-QSAR: AI inside a platform for predictive toxicology

Proceedings of the workshop "Popularize Artificial Intelligence 2013", December 5th 2013, Turin, Italy

Published on CEUR Workshop Proceedings Vol-1107

2.8.Availability of information about the model:

The model is non-proprietary and the training set is available.

2.9.Availability of another QMRF for exactly the same model:

Another QMRF is not available.

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Rodents

3.2.Endpoint:

TOX 7.7. Carcinogenicity (in vivo)**3.3.Comment on endpoint:**

Carcinogenicity is a very complex biochemical phenomenon involving processes at the cellular level. The carcinogenicity of a substance depends on its molecular structure and a certain number of phenomena which are only partially known. Typically, the carcinogenic process involves one or more processes, showing a relationship with the mutagenic potential of a substance, but other processes are possible for carcinogens which are non mutagenic

3.4.Endpoint units:

Adimensional

3.5.Dependent variable:

The dependent variable is cancerogenic effect on rat, as binary classification: 0 (non-carcinogen), 1 (carcinogen)

3.6.Experimental protocol:

The rules (structural alerts) have been extracted with SARpy software from a dataset obtained from the union of the ISS carcinogenicity (ISSCAN) database from Istituto Superiore della Sanità [1], and of the Carcinogenicity Genotoxicity eXperience (CGX) dataset [2]

3.7.Endpoint data quality and variability:

Carcinogenicity data on rodents have been processed by human experts (from ISS and JRC). In more detail: Experimental carcinogenicity data and chemical structures included in the ISSCAN database have been curated by ISS scientists (toxicological data assessed and critically selected). The EURL ECVAM Genotoxicity & Carcinogenicity Consolidated database is a structured master database that compiles available genotoxicity and carcinogenicity data for Ames positive chemicals originating from different sources, including regulatory agencies, industry and literature databases covering different sectors (e.g., US-NTP, EFSA, SCCS, Cosmetic Europe, BASF, ECHA, ISSTox, ...). Only chemicals with a known chemical identity (structure, purity, molecular weight, CAS number) and valid in vitro and in vivo results for the genotoxicity endpoints and/or for carcinogenicity were included. "Overall Calls" were defined for each genotoxicity assay in vitro and in vivo and carcinogenicity by following defined criteria for the reliability of each study and quality of data for those chemicals appearing in more than one source with different calls. 4 categories were considered (+), (-), (E) and (I). Where information was missing, even for those chemicals with one single data entry, scientific literature was consulted

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Expert rule-based system

4.2.Explicit algorithm:

Expert rule-based system

Set of 43 rules (structural alerts) related to carcinogenic activity. These rules are expressed SMARTS representing molecular fragments (reported in section 4.3). If at least one rule is matching with the target compound, a “Carcinogen” prediction is given. Otherwise, a “Possible NON-Carcinogen” prediction is given

4.3.Descriptors in the model:

[1]O=NNCC

[2]c1occc1

[3]O=CN(N)C

[4]CCCN(CC)CC

[5]C1CC(=CC)CCC1

[6]Nc1ccc(cc1C)C[7]NCCCN

[8]O=S(=O)(OC)

[9]c1ccc2OCOc2c1

[10]Nc1ncccc1

[11]N(CCCl)CCCl

[12]c1cn(cnc1)

[13]C=C(C=C)C

[14]O=NNC

[15]O=P(OC)

[16]O(c1ccc(cc1)CC=C)

[17]c1ncn(c1)C

[18]C(CCCC(CC)Cl)Cl

[19]c1ncsc1

[20]C=CCN

[21]O=Cc1ccccc1O

[22]O(c1ccc(cc1N))C

[23]O1CC1C

[24]SN(C)C

[25]C(CCl)Cl

[26]c1c(cc(cc1Cl)Cl)Cl

[27]NNCC

[28]O=CN(N)

[29]C(OC)C(C)C

[30]c1ccc2cc(ccc2c1)

[31]Nc1cccc(c1C)C

[32]NNc1ccccc1

[33]c1cc(ccc1C)Cl

[34]N(CCO)CCO

[35]Nc1ccc(cc1N)

[36]c1ccc(cc1N)C

[37]O(c1ccc(cc1)C)C

[38]C(c1ccccc1)CO

[39]C(=CCC)CC

[40]N(Cc1ccccc1)C

[41]Nc1ccc(cc1)C

[42]Nc1ccccc1

[43]n1cccc(c1)

4.4.Descriptor selection:

The SARpy software has been used with a cross-validated procedure, ending with the extraction of a set of 43 rules (structural alerts) related to carcinogenic activity

4.5.Algorithm and descriptor generation:

The 43 rules (structural alerts) are expressed as SMARTS representing molecular fragments. The SARpy software was used to extract the rules from the two carcinogenicity datasets. SARpy breaks the chemical structures of the compounds in the training set into fragments of a desired size, and it identifies fragments related to the target property. It then also shows the fragments related to the effect. Inhibiting conditions are identified which prevent the appearance of the effect, even in presence of the active fragment. The system uses SMILES in the canonical form. It allows choice in building more conservative or more accurate models

4.6.Software name and version for descriptor generation:

SARpy software free tool to develop a model to classify chemicals according to a given property Giuseppina Gini, Politecnico di Milano (giuseppina.gini@polimi.it); Emilio Benfenati, Istituto di Ricerche Farmacologiche Mario Negri (emilio.benfenati@marionegri.it)http://sarpy.sourceforge.ne

4.7.Chemicals/Descriptors ratio:

Not applicable to expert systems

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The Applicability Domain (AD) is assessed using the original algorithm implemented within VEGA. An overall AD index is calculated, based on a number of parameters, which relate to the results obtained on similar chemicals within the training and test sets.

If 1 ≥ AD index ≥ 0.8, the predicted substance is into the Applicability Domain of the model. It corresponds to good reliability of prediction.

If 0.8 > AD index ≥ 0.6, the predicted substance could be out of the Applicability Domain of the model. It corresponds to moderate reliability of prediction.

If AD index < 0.6, the predicted substance is out of the Applicability Domain of the model and corresponds to low reliability of prediction

5.2.Method used to assess the applicability domain:

The Applicability Domain and chemical similarity is measured with the algorithm developed for VEGA. Full details are in the VEGA website (www.vegahub.eu), including the open access paper describing it [3]. The AD also evaluates the correctness of the prediction on similar compounds (accuracy), the consistency between the predicted value for the target compound and the experimental values of the similar compounds, the range of the descriptors, and the presence of unusual fragments, using atom centred fragments.

Similar molecules with known experimental value:

This index takes into account how similar are the first two most similar compounds found. Values near 1 mean that the predicted compound is well represented in the dataset used to build the model, otherwise the prediction could be an extrapolation. Defined intervals are:

If 1 ≥ index > 0.80, strongly similar compounds with known experimental value in the training set have been found.

If 0.80 ≥ index > 0.6, only moderately similar compounds with known experimental value in the training set have been found.

If index ≤ 0.6, no similar compounds with known experimental value in the training set have been found.

Accuracy (average error) of prediction for similar molecules:

This index takes into account the classification accuracy in prediction for the two most similar compounds found. Values near 1 mean that the predicted compounds fall in an area of the model's space where the model gives reliable predictions (no misclassifications), otherwise the lower is the value, the worse the model behaves. Defined intervals are:

If 1 ≥ index > 0.90, accuracy of prediction for similar molecules found in the training set is good

If 0.9 ≥ index > 0.5, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≤ 0.5, accuracy of prediction for similar molecules found in the training set is not adequate

Concordance for similar molecules:

This index takes into account the difference between the predicted value and the experimental values of the two most similar compounds. Values near 0 mean that the prediction made disagrees with the values found in the model's space, thus the prediction could be unreliable. Defined intervals are:

If 1 ≥ index > 0.90, molecules found in the training set have experimental values that agree with the target compound predicted value

If 0.9 ≥ index > 0.5,, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index ≤ 0.5, similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value

Atom Centered Fragments similarity check: This index takes into account the presence of one or more fragments that aren't found in the training set, or that are rare fragments. First order atom centered fragments from all molecules in the training set are calculated, then compared with the first order atom centered fragments from the predicted compound; then the index is calculated as following: a first index RARE takes into account rare fragments (those who occur less than three times in the training set), having value of 1 if no such fragments are found, 0.85 if up to 2 fragments are found, 0.7 if more than 2 fragments are found; a second index NOTFOUND takes into account not found fragments, having value of 1 if no such fragments are found, 0.6 if a fragments is found, 0.4 if more than 1 fragment is found. Then, the final index is given as the product RARE * NOTFOUND. Defined intervals are:

If index = 1, all atom centered fragment of the compound have been found in the compounds of the training set

If 1 > index ≥ 0.7, some atom centered fragment of the compound have not been found in the compounds of the training set or are rare fragments

If index < 0.7, a prominent number of atom centered fragments of the compound have not been found in the compounds of the training set or are rare fragments

Model descriptors range check:

This index checks if the descriptors calculated for the predicted compound are inside the range of descriptors of the training and test set. The index has value 1 if all descriptors are inside the range, 0 if at least one descriptor is out of the range. Defined intervals are:

Index = TRUE, descriptors for this compound have values inside the descriptor range of the compounds of the training set

Index = FALSE, descriptors for this compound have values outside the descriptor range of the compounds of the training set

5.3.Software name and version for applicability domain assessment:

VEGA

Included in the VEGA software and automatically displayed when running the model

emilio.benfenati@marionegri.it

https://www.vegahub.eu/

5.4.Limits of applicability:

The model is not applicable to inorganic chemicals and substances containing unusual elements (i.e., different from C, O, N, S, P, Cl, Br, F, I). Salts can be predicted only if converted to the neutralized form.

6.Internal validation - OECD Principle 4

6.1.Availability of the training set:

Yes

6.2.Available information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: Yes

INChI: No

MOL file: No

NanoMaterial: No

6.3.Data for each descriptor variable for the training set:

No

6.4.Data for the dependent variable for the training set:

All

6.5.Other information about the training set:

The combined training set used for the extraction of the rules and model validation consists of 985compounds (733 carcinogens,252 non-carcinogens)

6.6.Pre-processing of data before modelling:

NA

6.7.Statistics for goodness-of-fit:

Training set: n = 985, Accuracy = 0.73; Specificity = 0.60; Sensitivity = 0.78, MCC 0.35.

TP 569, TN 150, FP 102, FN 164

6.8.Robustness - Statistics obtained by leave-one-out cross-validation:

NA

6.9.Robustness - Statistics obtained by leave-many-out cross-validation:

NA

6.10.Robustness - Statistics obtained by Y-scrambling:

NA

6.11.Robustness - Statistics obtained by bootstrap:

NA

6.12.Robustness - Statistics obtained by other methods:

Five-fold cross-validation:

Accuracy = 73%; Sensitivity = 77%; Specificity = 41%. TP = 562/735; TN = 157/254; FP = 95/254; FN = 172/735. MCC = 0.36

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

Yes

7.2.Available information for the external validation set:

NA

7.3.Data for each descriptor variable for the external validation set:

No

7.4.Data for the dependent variable for the external validation set:

No

7.5.Other information about the external validation set:

The predictability of the model has been evaluated on ECHA dataset, carcinogenicity data collected from the eChemPortal inventory (258 compounds) [3].

7.6.Experimental design of test set:

NA

7.7.Predictivity - Statistics obtained by external validation:

External validation on ECHA dataset:

Accuracy = 64%; Sensitivity = 48%; Specificity = 72%; TP = 43/89; TN =121/169; FP = 48169; FN = 46/89; MCC = 0.20

7.8.Predictivity - Assessment of the external validation set:

NA

7.9.Comments on the external validation of the model:

Accuracy, sensitivity, specificity, and the MCC for the external evaluation are determined using SARpy. Although the external evaluation is considered the best mean for the assessment of the predictive ability of a (Q)SAR model, the results of the external evaluation of any model are highly related to the relative similarity of the external evaluation set in relation to the training set

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The extracted 43 rules ("active" fragments, or structural alerts)represent structural fragments associated to carcinogenic acitivity.Inhibiting conditions are identifiedwhich prevent the appearance of the effect, even in presence of the active fragment

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori: rules (i.e., "active" fragments, or structural alerts) areextracted by SARpy software and not defined a priori

8.3.Other information about the mechanistic interpretation:

Additional information on fragments analysis are provided in the original publication [3]

9.Miscellaneous information

9.1.Comments:

NA

9.2.Bibliography:

[1]Benigni R, Battistelli CL, Bossa C, Tcheremenskaia O, Crettaz P (2013) New perspectives intoxicological information management, and the role of ISSTOX databases in assessing chemicalmutagenicity and carcinogenicity. Mutagenesis 28 (4), 401–409. doi:10.1093/mutage/get016https://academic.oup.com/mutage/article/28/4/401/2459896

[2]A. Golbamaki, E. Benfenati, N. Golbamaki, A. Manganaro, E. Merdivan, A. Roncaglioni, G. Gini(2016) New clues on carcinogenicity-related substructures derived from mining two large datasets ofchemical compounds. JOURNAL OF ENVIRONMENTAL SCIENCE AND HEALTH, PART C, VOL. 34, NO. 2, 97-113 http://dx.doi.org/10.1080/10590501.2016.1166879

[3] Floris, M., Manganaro, A., Nicolotti, O. et al. A generalizable definition of chemical similarity for read-across. J Cheminform 6, 39 (2014). https://doi.org/10.1186/s13321-014-0039-1

9.3.Supporting information:

Training set(s)Test set(s)Supporting information:

All available dataset are present in the model inside the VEGA software.

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

To be entered by JRC

10.2.Publication date:

To be entered by JRC

10.3.Keywords:

To be entered by JRC

10.4.Comments:

To be entered by JRC